Supplementary Material

For “Advancing functional connectivity research from association to causation”

1. Applying the FC mechanism framework to common FC measures

1.1. Pearson correlation with resting-state fMRI

In this section we will apply the framework to a commonly-used FC measure: Pearson correlation with resting-state fMRI[1,2]. There are some aspects of the example that will generalize to all uses of Pearson correlation with resting-state fMRI, while other aspects will be specific to this particular example dataset and analysis strategy. Notably, in this and the following two sections, we are only highlighting examples of how the proposed framework might be applied to specific FC approaches. These are not meant to be definitive, prescriptive, or exhaustive; on the contrary, an important aspect of this framework is that it should flexibly account for the changing landscape of neuroscientific methods and knowledge.

Step 1: Identifying target theoretical properties

In this example, we are primarily interested in characterizing neural interactions among a predefined set of brain regions during resting state. For simplicity, we restrict our inferences to linear causal interactions, quantified as Pearson correlations. One major advantage of using Pearson correlations is that this measure is well understood relative to alternatives, thus improving confidence in understanding its strengths and weaknesses as a measure of FC theoretical targets of interest. As an example of a well-understood weakness of this measure, we must accept that a third region may be acting to modulate (or fully cause) a correlation between any two regions, since Pearson correlation only models time series relationships in pairs. Thus, our target inferences are necessarily ambiguous (**Supplementary Table 1**), with additional ambiguities due to the poor temporal resolution of fMRI, both in terms of the whole-brain volume acquisition time (TR; here 785 ms) and the temporally-extended hemodynamic response function. We can therefore infer a causal interaction exists in the network, but that it is ambiguous with respect to directness and directionality. Moreover, while the presence of significant correlations indicates that causal edges exist in the network, our inference is also ambiguous with respect to which edge is causal (specificity). Nonetheless, the correlation-based network structure provides information that allows us to form probabilistic inferences about specific edges.

Step 2: Identifying methodological properties

While it would be theoretically ideal to measure interactions among individual neurons, fMRI only allows for measurement of large neural populations. **Supplementary Table 1** lists this and other methodological properties for this example study. Here we will use Pearson correlations to assess interactions between regions. One major reason for this decision is that efforts to rule out neural modulation from third regions (e.g., partial correlation) often lead to over-sparsification of FC graphs, collapsing an interaction among multiple neural units down to a single pairwise relationship (but see **Box 2**). **Supplementary Table 1** also lists the key assumptions required for our inferences to be

valid. Describing the observational pathway can be considered part of the methodological properties step, since methodological properties map observations to target theoretical properties. In this case, we ultimately seek to link observed resting-state FC with fMRI to action potentials (APs), given that we are interested in inferences about causal interactions among neural populations. Note that we clearly acknowledge weaknesses and ambiguities in our inferences regarding causal interactions and APs in several locations within **Supplementary Table 1**.

Step 3: Identifying confounding properties

Factors that can invalidate the target theoretical inferences due to alternative causal chains are listed here. There are a variety of potential confounders for fMRI, such as respiration and heart rate changes. These confounders become more problematic for resting-state FC analyses (relative to standard fMRI task activation analyses) given that FC measures typically characterize covariance regardless of whether it is time-locked to experimentally-manipulated events. This reduces experimental control, allowing confounding variables more opportunity to correlate with variables of theoretical interest. **Supplementary Table 1** lists a variety of confounders that have been documented in the resting-state FC fMRI literature, along with strategies to reduce their influence on target inferences. Once all other sections of the table are filled out, the "Relevant inferences" section can be completed, concisely describing the valid inferences of interest that can be made using this FC method.

Supplementary Table 1 – Example of using standard template table summarizing a functional connectivity (FC) measure in terms of the framework. Many details are specific to this example, which is modeled after a particular study, rather than generalizing to all uses of this method. Also note that these properties can change over time as the field learns more about relevant underlying processes. * = Likely does not fully correct confound.

Functional connectivity measure properties (as applied in this study)

Pearson correlation with resting-state fMRI

Target theoretical properties

- Direct and/or indirect interactions (ambiguous)
- Undirected (ambiguous)
- Common interaction source(s) possible (ambiguous)
- Weighted by statistical certainty of non-zero interaction (variance shared relative to total variance)
- Linear interactions

Assumptions

- Causal interactions are due to action potentials propagated via axons influencing synaptic activity

Methodological properties

- Equipment: Siemens 3T Tim TRIO MRI
- Spatial resolution: 2 mm cubic, downsampled (averaging) into functional parcels[3]
- Spatial coverage: Whole-brain, 210 mm field-of-view; only cortex analyzed
- Temporal resolution: 0.785 s, filtered through hemodynamic response (~6 s to peak)
- Transformation to normal distribution: Fisher's z-transform (on Pearson correlations)
- Observation equation: Balloon-Windkessel model
- Interaction estimate: Pearson correlation coefficient
- Neural entity: Brain regions (via voxel averaging)

See main text for more details

Assumptions

- Causal link to action potentials:
  Blood-oxygenation-level-dependent (BOLD) signal driven by synaptic activity (which is driven by action potentials); detected by effect on magnetic field
- Homogeneity of the hemodynamic response across regions
- Statistical assumptions for one-sample t-test at group level (inter-subject variance of Fisher's z-transformed Pearson correlations): normality (robust to small violations), independence, no outliers

Confounding properties

- Subject movement during scan influences correlations[4,5]
  - *Correction strategy*: Scrubbing/censoring[6], regression with motion parameters, CompCor[7,8] (regression with principal components of white matter, ventricle data), ICA-FIX[9]
- Respiration and heart rate induces global changes in oxygenation and blood flow, influencing correlations[10]
  - *Correction strategy**: Regression with mean white matter and ventricle signals, CompCor[7,8]
- Spatial autocorrelation of BOLD signal, leading to inflated local correlations
  - *Correction strategy*: Use of predefined parcels/regions

See main text for more details

Observational pathway

APs → synaptic activity/LFPs → HRF [CBV, Hb, CBF] → MR BOLD contrast

Relevant inferences

- Given a statistically significant group-level Pearson correlation between two regions' time series, and to the extent that the confound correction strategies were effective:
  The two regions interact (directly or indirectly, with ambiguous directionality), and/or are similarly influenced by common region(s), during resting state.

1.2. Task-based EEG with source localization and Granger-type influence

Step 1: Identifying target theoretical properties

Our benchmark here is a simple visual object recognition task and a high-density (128-channel) EEG recording, including inferior electrode placements (**Supplementary Table 2**). As for the preceding example, how well we can estimate the target theoretical properties will be limited by both physical limitations inherent to the forward problem and the properties of the methodological choices. We illustrate using a directed measure of source FC to assess the directionality of information/activity transfer, as well as the strength (weight) of this interaction.

Our efforts will be limited by the nature of the EEG forward problem. Volume conduction mixes and spatially smears the activity of neural entities. As a consequence, inferring “real” neural FC among entities by simple interpolation of FC measures in so called “sensor space” is in general impossible. To improve inference we will estimate the activity of estimated neural entities by solving the inverse problem of the EEG. The method selected for this purpose is eLoreta which is based outputs the time series of the estimated neural entities or “sources”. By the very nature of this linear inverse solution, the estimated neural entities will be a spatially smooth solution, with a limited spatial resolution. This implies that, while source localization dramatically improves spatial resolution over raw sensors, this particular method produces spatially blurred estimates of the neural entities and therefore mixes the connectivity of the actual neural entities – a phenomenon known as “leakage”. Achieving super resolution would eliminate leakage but the purpose of our example is to show the limitations of the estimates which follow from our choices.

Supplementary Table 2 – Example of using standard template table to summarize a task-based EEG study. Many details are specific to this example. Also note that these properties can change over time as the field learns more about relevant underlying processes. * = Likely does not fully correct confound.

Functional connectivity measure properties (as applied in this study)

Task-related Granger-type influence with scalp EEG

Target theoretical properties

- Direct and indirect interactions (ambiguous)
- Directed interactions
- Weighted interactions
- Linear interactions

Assumptions

- Causal interactions are due to action potentials propagated via axons influencing post-synaptic potentials that produce currents reflected as voltage differences on the scalp
- We have a fairly accurate model of the effects of volume conduction

Methodological properties

- Task: Visual object recognition
- Equipment: 128-channel active electrodes
- Spatial resolution: Inter-sensor distance 2.7 cm; Modeled source spread 6-7 mm[11]
- Spatial coverage: Whole-brain, biased by dipole orientation, amplitude, superficiality
- Bandpass filter: 3-45 Hz
- Temporal resolution: 1 ms
- Observation equation: estimated with the boundary element method (BEM) from T1 MRI
- Source localization method: eLORETA[12]
- FC measure for the sources : Granger-type influence
- Achievable estimate of neural entity: current source density with limited spatial resolution

See main text for more details

Assumptions

- Activity of all relevant neural entities is reflected in voltage scalp measurements
- Source current sources are spatially smooth
- Statistical assumptions: normality, independence, absence of outliers, linearity
- Volume conductor model and spatial smoothness allow approximate estimation of sources

Confounding properties

- Physiological artifacts (respiration, cardiovascular, skin conductance, eye blinks/movements)
  - *Correction strategy**: Measure and regress out of time series
- Motion artifacts
  - *Correction strategy**: High-pass filter, manual artifact removal
- Electromagnetic artifacts
  - *Correction strategy*: Faraday cage, active electrodes, notch filter
- Bias for simultaneous sources: occlusion of weak or deep sources[13]
  - *Correction strategy**: None
- Spurious correlations due to imperfections in head model
  - *Correction strategy*: Use T1-based reconstruction (Freesurfer)

See main text for more details

Observational pathway

APs → PSPs → Primary current source density → Volume conduction → Scalp potential

Relevant inferences

- If tests for Granger-type influence are significant, there was a *directed* interaction from one neural entity to another during visual object recognition
- Magnitude of difference in measures derived from Granger-type model indicates the relative *weight* of the interaction between task conditions

Step 2: Identifying methodological properties

Our *observation equation* specifies how neural sources map to scalp electrode recordings. This is also known as the *forward problem* and its numerical estimation results in the so-called lead field, mapping current sources in neural entities to electrodes. We will use the boundary element method (BEM)[14] for this calculation, which requires an approximation of cortical geometry using anatomical MRI[15,16]. The estimation of the neural entities, summarized by current source densities, is known as the *inverse problem*. This activity in each neural entity is approximated by a vertex on the cortical surface. Here we will use exact low resolution brain electromagnetic tomography (eLORETA)[12].

Our approach entails some important assumptions: (1) the activity of a neural entity is reflected in electrode voltages; (2) current sources are spatially smooth; (3) the volume conduction model is accurate; and (4) the head model is accurate[17].

Given the target inferences defined above, our FC estimate will be Granger-type influence[18,19], which is a directed measure that provides weighted estimates of connectivity for specific frequency bands. To stress the difference with what we refer to here with "causal" we deliberately omitted the often used, and often misinterpreted "Granger causality" name.

Step 3. Identifying confounding properties

EEG is highly susceptible to environmental and physiological artifacts. These include line noise generated by electrical circuitry, artifacts generated by ocular muscles, subject motion, respiration, and cardiovascular events. Several measures can be taken to address these potential confounds. Environmental noise can be filtered using a Faraday cage, or by using active electrodes[20]. Other sources of noise can be removed by band-pass or notch filtering approaches. Ocular artifacts can be measured directly via ocular electrodes or simultaneous eye tracking[21].

The eLORETA approach also has at least two major limitations that may produce imperfect estimates of the activity of neural entities. First, it has been found to miss sources if they are deep or low-amplitude, and occluded by simultaneously active superficial or higher-amplitude sources[13]. This affects spatial coverage by biasing the set of sources most likely to be attributed to a given pattern of scalp electrode activity. Second, this approach suffers from "leakage" between individual sources, resulting from a failure to sufficiently unmix the activity of neural entities signals[22].

1.3. Cross-correlation of spike trains (invasive recording)

Here we consider a hypothetical study in which single neurons will be recorded in awake behaving rats (**Supplementary Table 3**). Interactions between neurons recorded extracellularly will be investigated by measuring correlations in their spiking across a range of time lags, yielding a cross-correlation function. Despite the breadth of this topic, much of it is encompassed by the mechanistic FC framework described here.

Step 1: Identifying target theoretical properties

The original goal of cross-correlation analysis was to identify synaptic connections between neurons whose spike trains were simultaneously recorded[23]. Depending on the lag in the maximal deviation of the cross-correlation function and whether it is positive or negative with respect to baseline, one can infer either an excitatory or inhibitory synapse, its direction, and strength[23–25]. However, the massive convergence and divergence of connections in cortical circuits, and weakness of any particular connection between neurons, often leads to false negatives[26]. Instead, aggregate

interactions between populations of neurons (rather than single cells) will be analyzed in this study, which typically manifest as zero-time lag (i.e., zero-phase lag) peaks in the cross-correlation function[27]. Zero-time lag correlations are associated with numerous cognitive processes[28], highlighting their relevance to neurocognitive function.

In principle, a peak in the cross-correlation function between two neurons could arise from a shared third afferent source[29–31], as is the case with other FC measures. We will address this source of ambiguity using a spike-triggered joint histogram, which measures the interaction between spike trains from multiple neurons[32], allowing us to account for the spiking of an afferent structure. Our target theoretical property is thus an indirect connection between recorded neurons that is not mediated by a third source.

Step 2: Identifying methodological properties

Here, spike trains from single neurons (our neural entities) will be recorded extracellularly by means of fine tipped electrodes, and the resulting time series will be subjected to cross-correlation analysis. When neurons within ~100 μm of the electrode tip fire action potentials they produce a transmembrane current that, via coulombs law, can be detected as a change in voltage at the electrode/brain interface[33]. If both are embedded in the same recurrently connected network they can synchronize their spiking. We will use a fast sampling rate (30 kHz) to capture spiking from individual neurons, and multichannel recordings that allow for sorting of spikes either into groups of multiple units or well-isolated single units.

Step 3: Identifying confounding properties

Because our target property is an indirect interaction not driven by a common extrinsic source, we have to adjust our analyses for potential confounds. To rule out comodulations driven exclusively by task events, there are a variety of permutation techniques available[34]. We will also use a jittering approach, with a jitter window of 1000 ms, to control for slow state dependent changes in spiking. Finally, when neurons have low firing rates, there is a risk that false negatives will occur, in which case activity may have to be sampled for a longer time[26].

Supplementary Table 3 – Example of using standard template table to summarize an invasive spike train analysis. Many details are specific to this example. Also note that these properties can change over time as the field learns more about relevant underlying processes.

Functional connectivity measure properties (as applied in this study)

Cross-correlation of spike trains

Target theoretical properties

- Direct (monosynaptic) and indirect (polysynaptic/comodulation)
- Directed (based on temporal lag)
- Linear interactions
- Weighted by interaction strength

Assumptions

- Causal interactions are due to action potentials propagated via axons influencing downstream spike probability

Methodological properties

- Equipment: Metal electrodes, AC extracellular amplifiers
- Spatial resolution: ~100 μm
- Spatial coverage: From local microcircuits to between brain regions
- Temporal resolution: ~30 μs (sampling rate 30 kHz)
- Discretization of spike trains: Spikes times are binned in steps between 100 μs and 2 ms
- Observation equation: N/A (direct recordings)
- Interaction estimate: Cross-correlations
- Neural entity: Single neurons

Assumptions

- Causal link to action potentials: Extracellular waveforms reflect transmembrane currents. Action potentials produce sub-millisecond inward currents that appear as brief negative potentials on nearby recording electrodes.
- Statistical assumptions for coincident spike counts: stationary poisson processes

Confounding properties

- Individual spikes recorded at the same site cannot be reliably assigned to particular neurons
  - *Correction strategy*: Neurons can be better distinguished using dense multi-contact electrodes that allow more variation in spike waveform shape between nearby neurons
- An absence of coincident spikes at zero time lag arises from spike waveform 'shadowing'
  - *Correction strategy*: Use of template matching algorithm for spike detection instead of traditional threshold trigger, ignore deviations from the null at the zero time lag
- Peaks in the cross-correlation arise from covariations spiking driven by state changes or delivery of stimuli
  - *Correction strategy*: Generation of null distributions for the cross-correlation using jittering and permutation approaches

See Methods for more details

Observational pathway

APs → extracellular fields → stereotyped waveform → coincident spike counts

Relevant inferences

- Given a statistically significant peak in cross-correlation function:
  Peaks that are not at zero-lag, of short duration (<~5 ms), and have low jitter arise from probable monosynaptic connections. The magnitude of the peak's time lag corresponds to the axonal and synaptic conduction delay, while the sign of the lag is the direction of the interaction.

Supplemental references

1. Fox, M. D. & Raichle, M. E. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* **8**, 700–711 (2007).
2. Biswal, B., Yetkin, F. Z., Haughton, V. M. & Hyde, J. S. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.* **34**, 537–541 (1995).
3. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
4. Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* **59**, 2142–2154 (2012).
5. Van Dijk, K. R. A., Sabuncu, M. R. & Buckner, R. L. The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage* **59**, 431–438 (2012).
6. Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* **59**, 2142–2154 (2012).
7. Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage* **37**, 90–101 (2007).
8. Power, J. D. *et al.* Ridding fMRI data of motion-related influences: Removal of signals with distinct spatial and physical bases in multiecho data. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2105–E2114 (2018).
9. Salimi-Khorshidi, G. *et al.* Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage* **90**, 449–468 (2014).
10. Birn, R. M., Diamond, J. B., Smith, M. A. & Bandettini, P. A. Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. *Neuroimage* **31**, 1536–1548 (2006).
11. Song, J. *et al.* EEG source localization: Sensor density and head surface coverage. *J. Neurosci. Methods* **256**, 9–21 (2015).
12. Pascual-Marqui, R. D. Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods Find. Exp. Clin. Pharmacol.* **24 Suppl D**, 5–12 (2002).
13. Wagner, M., Fuchs, M. & Kastner, J. Evaluation of sLORETA in the Presence of Noise and Multiple Sources. *Brain Topogr.* **16**, 277–280 (2003).

14. Oostendorp, T. F. & van Oosterom, A. Source parameter estimation in inhomogeneous volume conductors of arbitrary shape. *IEEE Trans. Biomed. Eng.* **36**, 382–391 (1989).
15. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* **9**, 179–194 (1999).
16. Kim, J. S. *et al.* Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. *Neuroimage* **27**, 210–221 (2005).
17. Cho, J.-H., Vorwerk, J., Wolters, C. H. & Knösche, T. R. Influence of the head model on EEG and MEG source connectivity analyses. *Neuroimage* **110**, 60–77 (2015).
18. Nolte, G. *et al.* Robustly estimating the flow direction of information in complex physical systems. *Phys. Rev. Lett.* **100**, 234101 (2008).
19. Haufe, S., Nikulin, V. V., Müller, K.-R. & Nolte, G. A critical assessment of connectivity measures for EEG data: a simulation study. *Neuroimage* **64**, 120–133 (2013).
20. Laszlo, S., Ruiz-Blondet, M., Khalifian, N., Chu, F. & Jin, Z. A direct comparison of active and passive amplification electrodes in the same amplifier system. *J. Neurosci. Methods* **235**, 298–307 (2014).
21. Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M. & Kliegl, R. Coregistration of eye movements and EEG in natural reading: analyses and review. *J. Exp. Psychol. Gen.* **140**, 552–572 (2011).
22. Colclough, G. L., Brookes, M. J., Smith, S. M. & Woolrich, M. W. A symmetric multivariate leakage correction for MEG connectomes. *Neuroimage* **117**, 439–448 (2015).
23. Moore, G. P., Segundo, J. P., Perkel, D. H. & Levitan, H. Statistical signs of synaptic interaction in neurons. *Biophys. J.* **10**, 876–900 (1970).
24. Cope, T. C., Fetz, E. E. & Matsumura, M. Cross-correlation assessment of synaptic strength of single Ia fibre connections with triceps surae motoneurones in cats. *J. Physiol.* **390**, 161–188 (1987).
25. English, D. F. *et al.* Pyramidal Cell-Interneuron Circuit Architecture and Dynamics in Hippocampal Networks. *Neuron* **96**, 505–520.e7 (2017).
26. Schwindel, C. D., Ali, K., McNaughton, B. L. & Tatsuno, M. Long-term recordings improve the detection of weak excitatory-excitatory connections in rat prefrontal cortex. *J. Neurosci.* **34**, 5454–5467 (2014).
27. Martin, A. B. & von der Heydt, R. Spike synchrony reveals emergence of proto-objects in visual cortex. *J. Neurosci.* **35**, 6860–6870 (2015).

28. Uhlhaas, P. J. *et al.* Neural synchrony in cortical networks: history, concept and current status. *Front. Integr. Neurosci.* **3**, 17 (2009).

29. Brody, C. D. Correlations without synchrony. *Neural Comput.* **11**, 1537–1551 (1999).

30. Perkel, D. H., Gerstein, G. L. & Moore, G. P. Neuronal spike trains and stochastic point processes. II. Simultaneous spike trains. *Biophys. J.* **7**, 419–440 (1967).

31. Gerstein, G. L. & Clark, W. A. Simultaneous Studies of Firing Patterns in Several Neurons. *Science* **143**, 1325–1327 (1964).

32. Paz, R., Bauer, E. P. & Paré, D. Measuring correlations and interactions among four simultaneously recorded brain regions during learning. *J. Neurophysiol.* **101**, 2507–2515 (2009).

33. Buzsáki, G., Anastassiou, C. A. & Koch, C. The origin of extracellular fields and currents--EEG, ECoG, LFP and spikes. *Nat. Rev. Neurosci.* **13**, 407–420 (2012).

34. Grün, S. Data-driven significance estimation for precise spike correlation. *J. Neurophysiol.* **101**, 1126–1140 (2009).